CDD:006

A METHODOLOGY FOR MODELING THE COMPLEXITY OF THE HARTREE-FOCK PROCEDURE

METODOLOGIA PARA MODELAR A COMPLEXIDADE DO CÁLCULO HARTREE-FOCK

Jonathan de Matos[3], Eduardo Bortolato[1], Alexandre Camilo Jr.[2,*], Paulo S. L. de Souza[3]

State University of Ponta Grossa, Department of Computing[1] – Department of Physics[2]
[*]Corresponding author
E-mails: jonathandematos@gmail.com, eborto@gmail.com, acamilo@uepg.br
Av. Carlos Cavalcanti, 4748, 84030-900 – Ponta Grossa – PR – Brazil

University of São Paulo – USP[3] Institute of Mathematical Sciences and Computing - ICMC
Computer Science and Statistics Department – SCE
E-mail: pssouza@icmc.usp.br
Av. Trabalhador São-carlense, 400 – Caixa Postal: 668
13560-970 - São Carlos - SP – Brazil

Recebido para publicação em 20/05/2008
Aceito para publicação em 30/06/2008

ABSTRACT

The maximum number of two electrons integrals (2e- integrals) calculated in the Hartree-Fock (HF) method is given by N^4, where N is the number of basis functions involved in the calculation. However, in real situations, this amount of integrals can be reduced to the range $\sim N^{3.5}$ to $\sim N^2$, depending on factors such as: the molecular structure and the basis set used in the calculation. The methodology presented in this work allows for anticipating the real amount of 2e- integrals calculated in a HF procedure to different molecular structures. The proposal is based on the average of the inertia moments that represent the geometry of the executed molecule. The molecules have been divided in 3 groups of molecular geometry: 3D, planar and linear. The experiments considered molecules with regular and irregular geometries, in the STO-3G, 6-31G and 6-311G basis set. Calculations have been carried out using the GAMESS package. Results demonstrate a consistent behavior for the methodology proposed, as for molecules with regular geometry and for molecules with more irregular geometric structure. The results presented in this paper will allow one to estimate the demand for hard disk and CPU generated in the execution of a molecule with the HF procedure.

Keywords: GAMESS. HF Procedure. 2e- integrals cutoff prediction

RESUMO

O número máximo de integrais de dois elétrons (integrais de 2e-), calculado no método de Hartree-Fock (HF), é dado por N^4, em que N é o número de funções de base envolvido no cálculo. Contudo, em situações reais, esta quantidade de integrais pode ser reduzida para a faixa a $\sim N^{3.5}$ para $\sim N^2$, dependendo de algum fatores, tais como: a estrutura molecular e a base utilizada no cálculo. A metodologia apresentada neste trabalho permite antecipar a quantidade real de integrais de 2e- função integral obtidas em um cálculo HF para diferentes estruturas moleculares. A proposta é baseada na média dos momentos de inércia que representam a geometria da molécula. As moléculas foram divididas em 3 grupos de geometria molecular: 3D, planar e linear. Os experimentos consideram moléculas com geometria regular e irregular, nas bases STO -3G, 6-31G e 6-311G. Os cálculos foram feitos usando o pacote GAMESS. Nossos resultados demonstram um comportamento consistente para a metodologia proposta, tanto para as moléculas com geometria regular, quanto para as moléculas com geometria irregular. Nossos resultados, apresentados neste artigo, permitem estimar a demanda de disco rígido e CPU gerada na execução de um cálculo HF para uma molécula.

Palavras Chave: GAMESS. Procedimento HF. Predição do corte das integrais de 2 elétrons

1 Introduction

The objective of the *ab-initio* Quantum Chemistry algorithms is to perform highly accurate calculations at the lowest computational demand (STROUT, 1995). In this way, different algorithms for the HF procedure have been considered in literature (CHALLACOMBE, 1997; CHALLACOMBE, 2000; GAN, 2003; SCHWEGLER, 1996; SCHWEGLER, 1997; SCHWEGLER, 1999; SCHWEGLER, 2000; TYMCZAK, 2005a; TYMCZAK, 2005b). These proposals have the objective to reduce the complexity of the HF procedure, in a general sense.

Software tools applied to quantum chemistry constantly are improved and re-feed the theory of the area (TRUHLAR, 2000). The software improvements follow and also stimulate the constant technological advances of the computers. The role of the parallel computation is a common example of improvement in the current days and that makes possible executions forbidden before.

Besides producing correct algorithms, considering the Quantum Chemistry theory, the new proposals also attack the existing bottleneck in the HF procedures performance. These bottlenecks are mainly related with the CPU time and storage capacity. The generated demand of the computational devices (CPU, RAM memory, bus, hard disks and others) depends on factors as: (1) complexity of the calculation to be carried out (e.g.: molecular structure and basis set), (2) implementation of the method (e.g.: Direct SCF or Conventional SCF) and (3) computational system (e.g.: scalar or vectorial architecture, sequential or parallel machine, performance of the communication devices and hard disks).

A question remains open, despite of the constant improvements in the computation: how to foresee the computational cost to calculate the energy, in HF method, of determined molecular structure? If the final user will be able to answer this question, it will be possible to estimate, for example, if the available hard disks have capacity enough to store the 2e- integrals necessary to the simulation or if the CPU time selected for the execution is enough. Considering such questions is extremely important for the final user. It is very possible that the HF procedure executions occupy Gbytes of storage space and/or delay many hours/ days of CPU time.

The answer for the question above depends on different exclusive molecular parameters and, therefore, it is not trivial. The theoretical cost N^4 of the HF procedure is known in literature (ALMLÖF, 1982). Here N is the number of basis functions. This value should enable to estimate the amount of 2e- integrals to be calculated, activity that dominates the computational cost of the HF procedure (STROUT, 1995). However, it is also known that the theoretical cost N^4 is reduced in practical for $\sim N^{3.5}$ to $\sim N^2$, in the algorithm Direct SCF - DirSCF (ALMLÖF, 1982; SCHMIDT, 1993), depending on the molecular structure and basis set (STROUT, 1995).

In order to determine the computational cost of the calculation of the energy based on HF method it is necessary, therefore, to determine previously which is the real amount of 2e- integrals to be calculated. In other words, how many 2e- integrals will be discarded in the future execution of one determined molecular structure?

This work aims to estimate this real amount of integrals to be calculated in a HF procedure, considering determined standards of molecular structures and the algorithm Conventional SCF - CSCF in the GAMESS package (SCHMIDT, 1993). The geometry of the molecular structure to be simulated was considered to carry out this estimate. The metric used to represent geometry was the average of the molecular moment of inertia. The results demonstrate that this methodology is consistent for STO-3G, 6-31G and 6-311G basis set, thus it presents a trustworthy relation between molecular geometry and the amount of 2e- integrals discarded.

The methodology proposed in this work does not consider the effects of successive integrals screening realized in the iterations of the HF procedure; such as it is usual on DirSCF (SCHMIDT, 1993). The scope of this work is to determine the cutoff present when all 2e- integrals are previously evaluated, stored in disk and so they are used later.

This paper is organized as follow. In section 2 the scaling properties of the HF method are presented. Section 3 describes the methodology proposed in this work, using the average of the inertia moment to determine the amount of discarded 2e- integrals. Section 4 presents the results obtained with this methodology. Section 5 relates the conclusions and the perspectives of future works.

2 **Scaling Properties of the Hartree-Fock Method**

The Hartree-Fock method complexity was originally determined as N^4 due to (2.1)

$$(\mu\nu \mid \lambda\sigma) = \int\int \varphi_\mu(1)\varphi_\nu(1)\frac{1}{r_{12}}\varphi_\lambda(2)\varphi_\sigma(2)d\tau_1 d\tau_2 \qquad (2.1)$$

where $\mu, \nu, \lambda\, e\, \sigma$ are atomic orbitals (STROUT, 1995). However, a great part of the 2e- integrals has a negligible value. It allows the discarding of these values in the HF equations evaluation. This behavior is represented by the Schwarz inequalities given by (2.2)

$$|(\mu\mu \mid \lambda\sigma)| \leq \sqrt{(\mu\nu \mid \mu\nu)(\lambda\sigma \mid \lambda\sigma)} \qquad (2.2)$$

The appliance of this equation in HF calculations reduces the exponent scale N^α . As the size of the molecule increases, the reduction becomes more significant. Big molecular systems have a lot of atomic orbitals distant from each other. It causes less interaction among them (ALMLÖF, 1982; STROUT, 1995).

The influence of Schwarz inequalities also varies depending on the algorithm that implements uses the HF procedure. The impact is greater when these algorithms evaluate 2e- integrals in successive iterations, hence successive cuts in integrals amount are made. Schwarz inequalities gains are comparatively smaller when 2e- integrals are formerly evaluated, temporarily stored in disk and then used. This happens because integrals amount are discarded just on its evaluation, so the discarding occurs just once.

Different software tools apply these ideas. Gaussian (FRISCH, 2004), Dalton (HELGAKER, 2001), NWChem (KENDALL, 2000), Spartan

(SPARTAN), Turbomole (AHLRICHS, 1989) and GAMESS (SCHMIDT, 1993) are some examples of these tools.

Different authors have been studying the properties of scalability inherent to HF methods. Almlöf et al (ALMLÖF, 1982), in a pioneering work in 1982, used N^2 to study 2e- integrals behavior and explored integrals screening effects on HF methods for a series of nitrogen linear molecules up to 16 atoms. They observed the decreasing of 2e- integrals fraction as the molecule size increases. They also made a comparative study of three nitrogen isomers, each of them composed by 8 atoms (a linear molecule, a planar ring and a cubic isomer), observing that the fraction of 2e- integrals used is greater for the large dimensional order molecules (cubic isomer), following by planar ring and linear molecule.

Computers processing capacity has increased in last decades. Jointly, programs efficiency follows this increase due to calculation methodology improvements (AIKENS, 2004; ALEXEEV, 2002; BOLDING, 2000; CHOI, 2003; FAMULARI, 1998; FEDOROV, 2004; GAN, 2003; GLAESEMANN, 1998). In 1995, Strout and Scuseria (STROUT, 1995) presented details about effects of integrals screening on the scaling properties of HF methods in large molecular systems treated in (ALMLÖF, 1982). They have studied two molecular system models: one of them composed of graphitic sheets, of bi-dimensional features, and another one composed of diamond like three-dimensional structures. Graphitic sheets follow the homologous sequence $C_{6n}{}^{2}H_{6n}$ while diamond like structures follow the sequence $C_{\left(4\cdot n^3-n\right)/3}H_{4\cdot n^2}$. The structures have been studied using STO-3G, DZ e DZP. Authors verify 2e- integrals amount and CPU time for STO-3G basis set, when executing each molecular structure. These executions considered the first HF iteration.

They showed the scaling exponent behavior among molecule pairs used. This exponent, called here β , was defined as:

$$\left(\frac{N_2}{N_1}\right)^{\beta}=\left(\frac{I_2}{I_1}\right) \qquad (2.3)$$

what leads to the relation:

$$\beta=\frac{ln(I_2/I_1)}{ln(N_2/N_1)} \qquad (2.4)$$

where ***I*** is integrals number and ***N*** is basis functions number involved in a HF calculation for each molecule.

The most significant result is the scaling asymptotic behavior in the limit of large molecules. To graphitic sheets the obtained value was 2.1 and to diamond like structures it was 2.4. In all the cases, integrals screening based on integrals count reduced significantly the scaling exponent to a value close to ~2.

HF procedure has other time-consuming steps, besides 2e- integrals evaluation (STROUT, 1995). Fock matrix diagonalization is an example of a calculation done during HF procedure and scaling as N^3. (STROUT, 1995) have analyzed the proportionality of these time-consuming steps in order to prove the Fock matrix diagonalization contribution to HF procedure total wall clock. Their results showed that for large molecules the diagonalization time is less than two percent of total CPU time, while integrals calculation, in a consistent way, dominate the total time.

Results presented by (STROUT, 1995) show the importance of 2e- calculation in HF method. However, this study does not point out how to estimate the 2e- integrals cutoff scaling, taking in account different kinds of molecular structures and different basis set. In other words, there is no way to estimate the value of β without executing at least one HF cycle in order to determine 2e- integrals amount generated in the basis set combining with the molecular structure being simulated.

3 Using the Inertia Moment

Works described in the last section demonstrate that mathematical bounds computed with the Schwarz inequality screen and eliminate four-center two electron integrals smaller than a threshold (STROUT, 1995). However, it is not known how to quantify previously the real number of integrals with more precision, according to both number of basis function and peculiar molecule to be simulated. This is essential to anticipate the algorithm complexity, which is responsible to compute the Hartree-Fock energy in the GAMESS.

Table 1 shows a roll of distinct molecules and their respective 2e- integrals amount, considering 6-31 basis set. These molecules were grouped at: (a) three-dimensional, (b) planar and (c) linear. This classification follows the Almlöf et al. proposal (ALMLÖF, 1982), used by (STROUT, 1995) and allows the verifying that the amount of the eliminated 2e- integrals is actually higher for larger molecules (see last column). Third and fourth columns of the Table 1 show that the three-dimension group requires more integrals to be evaluated when compared to planar group, considering a proportional molecule size and a fixed screening threshold. The planar group itself requires more integrals than linear ones.

Table 1 – Relationship among distinct molecule types considering: maximal amount integrals, real amount integrals (in fact) evaluated and arithmetic-average of the inertia-moments X, Y and Z axis. Hartree-Fock results with an 6-31G basis set and a 10^{-10} hartree integral screening threshold. Table 1 (a) groups molecules with tri-dimensional structure; Table 1 (b) binds molecules with planar structure and Table 1 (c) join linear ones.

Table 1 (a)

Molecule	Cartesian	Integrals Amount (Peak)	Integrals Amount (Real)	Axis X Inertia Moment	Axis Y Inertia Moment	Axis Z Inertia Moment	Inertia Moments Average	Integrals Amount Index
C20	180	131220000	110959278	657.779	657.779	657.779	0.6578	0.8456
C24	216	272097792	208410000	930.973	930.976	1049.734	0.9706	0.7659
C26	234	374777442	273210000	1198.012	1198.009	1016.298	1.1374	0.7290
C32	288	859963392	546123744	1564.446	1784.801	1784.805	1.7114	0.6351
C36	324	1377495072	765225000	2297.974	2297.976	1909.063	2.1683	0.5555
C50	450	5125781250	1846036324	3976.07	3976.067	4571.637	4.1746	0.3601
C60	540	10628820000	2871443565	5938.251	5938.251	5938.251	5.9383	0.2702
C70	630	19691201250	3875575000	7421.579	8585.664	8594.906	8.2007	0.1968
C80	720	33592320000	4935675000	10796.679	10823.36	10866.172	10.8287	0.1469
tube9x0	810	53808401250	4784685000	13541.318	16706.462	16706.462	15.6514	0.0889
tube5x5	900	82012500000	6374865000	13943.243	21943.367	21943.372	19.2767	0.0777
tube10x0	900	82012500000	5480295000	20309.461	20309.461	18539.342	19.7194	0.0668
tube11x0	990	120075000000	6276855000	24464.076	24464.076	24640.614	24.5229	0.0523

Table 1 (b)

Molecule	Cartesian	Integrals Amount (Peak)	Integrals Amount (Real)	Axis X Inertia Moment	Axis Y Inertia Moment	Axis Z Inertia Moment	Inertia Moments Average	Integrals Amount Index
C024H12	240	414720000	91891108	1476.37	2952.745	1476.372	1.9685	0.2216
C054H18	522	9280941282	656138426	7425.74	14851.48	7425.74	9.9010	0.0707
C096H24	912	86474760192	2285019245	24008.771	24008.771	48017.543	32.0117	0.0264
C150H30	1233	288909830440	4307400000	45566.435	49412.205	94978.639	63.3191	0.0149
ST 1	141	49406770	12405000	174.8	1954.922	2129.722	1.4198	0.2511
ST 2	278	746602082	70680000	1667.401	7460.06	9127.461	6.0850	0.0947
ST 3	415	3707681328	187830000	7546.03	9454.102	17000.042	11.3334	0.0507
ST 4	552	11605565952	362790000	10114.626	19914.454	30029.079	20.0194	0.0313
ST 5	689	28170003480	588345000	15881.486	32119.028	48000.514	32.0003	0.0209
ST 6	826	58187567522	860745000	18376.166	55211.922	73582.089	49.0567	0.0148
ST 7	963	107501657770	1186680000	22646.283	84533.421	107179.704	71.4531	0.0110
ST 8	1100	183012500000	1552095000	25475.966	125582.391	151058.357	100.7056	0.0085

Table 1 (c)

Molecule	Cartesian	Integrals Amount (Peak)	Integrals Amount (Real)	Axis X Inertia Moment	Axis Y Inertia Moment	Axis Z Inertia Moment	Inertia Moments Average	Integrals Amount Index
PAH 5	226	326094722	53407147	378.47	4264.73	4643.21	3.0955	0.1638
PAH 6	266	625801442	77464180	451.36	7016.48	7467.84	4.9786	0.1238
PAH 7	306	1095962562	105564358	524.24	10750.46	11274.70	7.5165	0.0963
PAH 8	346	1791490082	139350678	588.39	15348.70	15937.09	10.6247	0.0778
PAH 9	386	2774976002	175805880	659.68	21387.37	22047.05	14.6980	0.0634
PAH 10	426	4116692322	216111579	730.98	28831.86	29562.84	19.7086	0.0525
Ppv02	150	63281250	14740941	202.10	1917.41	2119.51	1.4130	0.2329
Ppv03	234	374777442	41679123	278.21	7735.67	8013.88	5.3426	0.1112
Ppv04	318	1278257922	79897863	415.98	19742.05	20158.03	13.4387	0.0625
Ppv05	402	3264481602	130164851	522.92	40250.09	40773.02	27.1820	0.0399
Ppv06	486	6973568802	191671635	629.86	71492.38	72122.24	48.0815	0.0275
Ppv07	570	13195001250	264129359	736.80	115721.72	116458.53	77.6390	0.0200
Ppv08	654	22867622082	347291656	843.74	175190.93	176034.68	117.3565	0.0152
Ppv09	738	37079635842	440803634	950.69	252263.09	253213.77	168.8092	0.0119
Ppv10	822	57068608482	544969890	1057.62	348860.21	349917.83	233.2786	0.0095
Ppv11	906	84221467362	659156710	1164.56	467565.91	468730.47	312.4870	0.0078
Ppv12	990	1.20075E+11	783349636	1271.69	610531.35	611803.04	407.8687	0.0065

The values in third column (Integrals Amount Peak) are obtained from equation 3.1 (SCHMIDT, 1993),

$$\frac{N^4}{8} \qquad (3.1)$$

where $\boldsymbol{N}$ is the cartesian amount. The columns "Integrals Amount Real" and "Inertia Moments" (X, Y and Z) were obtained directly from GAMESS, by means of empirical experiments. The Integrals Amount Index is a normalized value (0 ≤ index ≤ 1) and it is determined from:

$$\frac{\textit{Integrals Amount Real}}{\textit{Integrals Amount Peak}} \qquad (3.2)$$

An index whose value is 0 represents a 100% *cutoff* and a number 1 represents that there was no cutoff somehow. This index can estimate the 2e- integrals real amount evaluated in fact for molecule, when applied to 2e- integrals maximum amount (from Eq. 3.1).

The starting point of this work was to analyze the molecular geometry, in order to consider the molecular spatial structure. Therefore, the arithmetic average $\boldsymbol{M}$ of the molecules moment of inertia X, Y and Z was joined to the 2e- integrals cutoff. Table 1 shows in its two last columns that the growth of the inertia moments average is proportional to the increase of the integrals that are removed from energy evaluation, when the three molecular groups are considered. This integrals cutoff growth is asymptotic, never exceeding the 100% limit.

Equations 3.3, 3.4 and 3.5 use the inertia moments average in order to determine the 2e- integrals cutoff. Each equation represents one of the three molecular groups cited previously, respectively: three-dimensional, planar and linear.

$$cut\_off_{3D} = \left(\frac{1}{0{,}861141 + 0{,}3578 \cdot M} \right) \qquad (3.3)$$

$$cut\_off_{planar} = \left(\frac{1}{01{,}97275 + 1{,}37505 \cdot M} \right) \qquad (3.4)$$

$$cut\_off_{linear} = \left(\frac{1}{2{,}92809 + 1{,}04832 \cdot M} \right) \qquad (3.5)$$

where $\boldsymbol{M}$ is arithmetic average from X, Y and Z inertia moments. All these three equations provide a normalized index, likewise made clear for Integrals Amount Index (last column of the Table 1). Again, this index approximates the 2e- integrals amount that will be evaluated for molecule in fact, when applied on 2e- integrals maximum amount (Eq. 3.1).

The equation model above is known as "inverse regression". It was defined empirically, comparing percentage of the 2e- integrals discarded to inertia moments average. This study considered the 6-31basis set.

$$\left(\frac{1}{(a + b \cdot M)} \right) + c \qquad (3.6)$$

a, *b* and *c* in 3.6 were defined for each molecular group, by means of iterative method for non-linear curve fitting. The GRACE software tool was used to obtain these coefficients (GRACE, 2009).

The three equations (3.3-3.5) are necessary to estimate the cutoff because the geometry influence causes specific absolute values inside each group, in despite of inertia moment be_consistent at the three molecular groups. Empirical results in Table 1 can show this feature.

4 Results and Discussion

Results described in this section demonstrate the modeling efficiency when compared to empirical (real) execution on GAMESS (SCHMIDT, 1993). Evaluations use the three molecular groups cited previously (three-dimensional, planar and linear) and also a fourth extra group. Molecules presented in the three first groups were chosen to allow the modeling scalability analysis, in face to gradual increase of molecular size. Fourth molecules group presents distinct-geometry features when comparing to other ones. This difference allows to show how ample the reach of this work is. The evaluations were done without symmetry and consider first the 6-31G basis set. STO-6G and 6-311G basis set were also considered in order to demonstrate the cutoff-modeling behavior with other basis set.

Three-dimensional group used the following molecules: fullerenes and nanotubes (Figs. 1 up to 4).

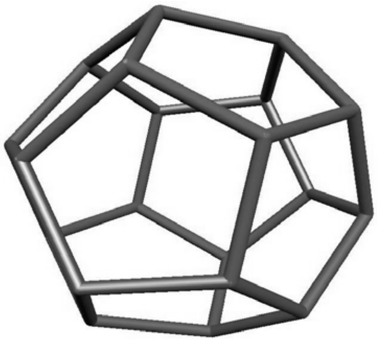

Figure 1 – Fullerene (C20 - 20 carbon atoms)

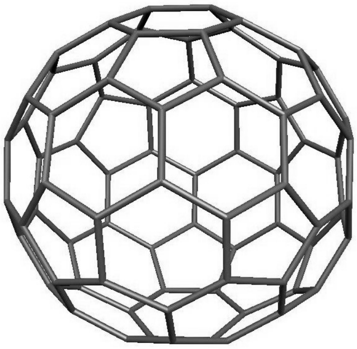

Figure 2 – Fullerene (C80 - 80 carbon atoms)

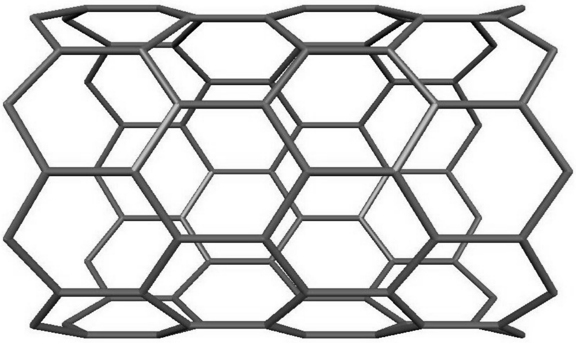

Figure 3 – Nanotube (5x5 – 100 carbon atoms)

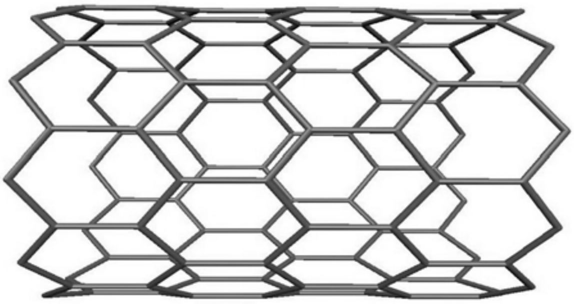

Figure 4 – Nanotube (11x0 – 100 carbon atoms)

Planar group used the molecules: graphitic sheets (Figs. 5 up to 6) and poly(3- β -steryl-thiophene) (ST)(Figs. 7).

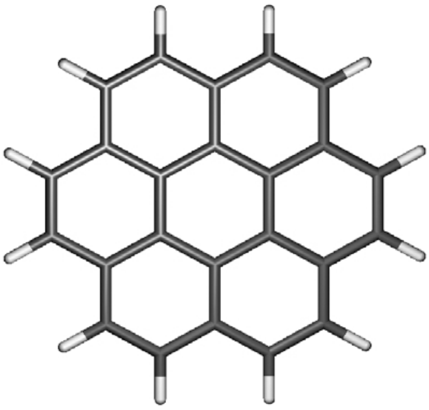

Figure 5 – C024H12 graphitic sheet

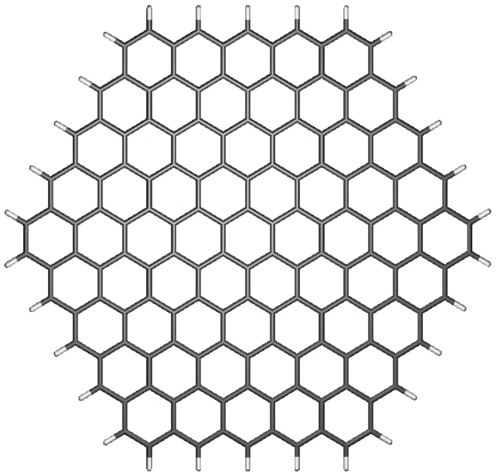

Figure 6 – C150H30 graphitic sheet

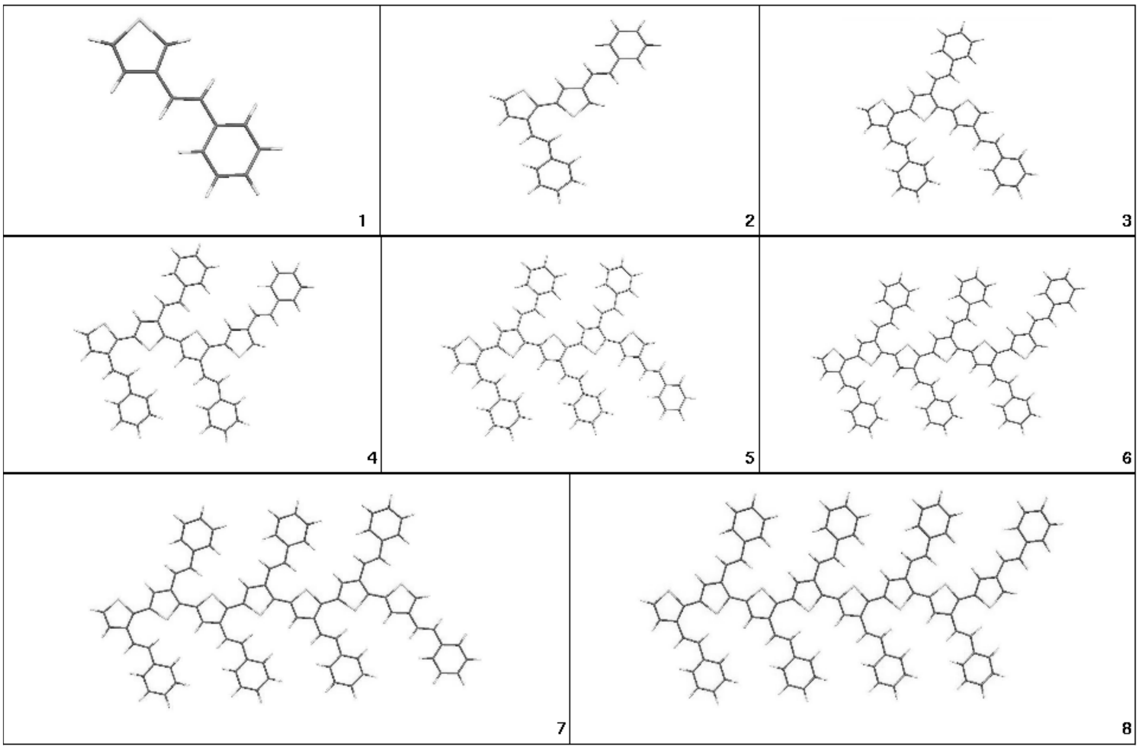

Figure 7 – poly(3- β -steryl-thiophene) (ST) model

Linear group considered the molecules: Polycyclic Aromatic Hydrocarbons (PAH's from 05 up to 10 units – Fig. 8) and a conjugated polymer: Poly-p-Phenylene Vinylene (PPV's from 02 up to 12 units – Fig. 9).

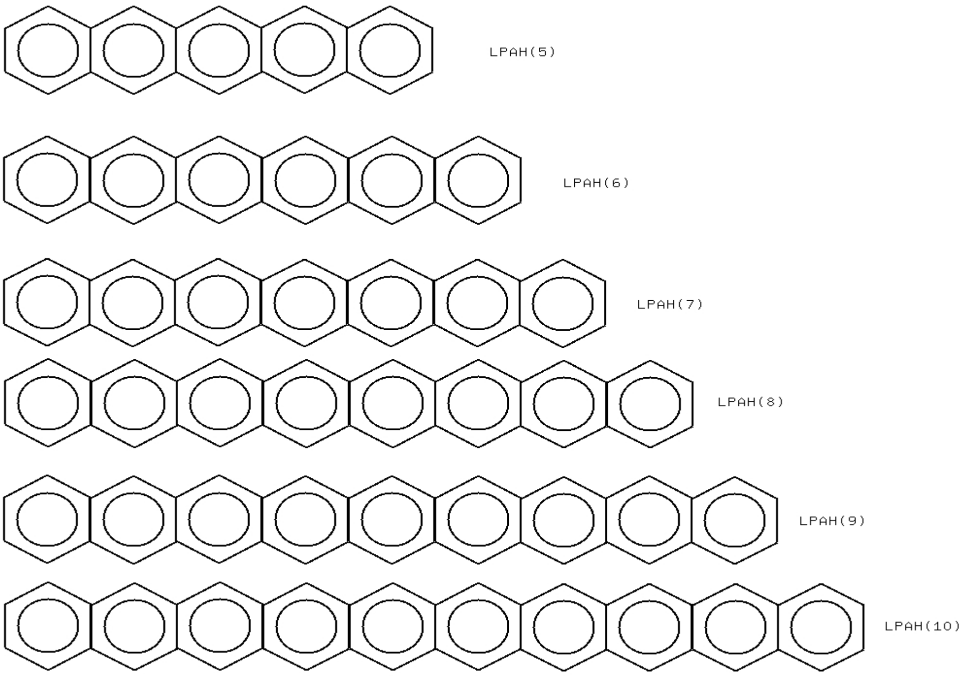

Figure 8 – Linear PAH molecules used in this work.

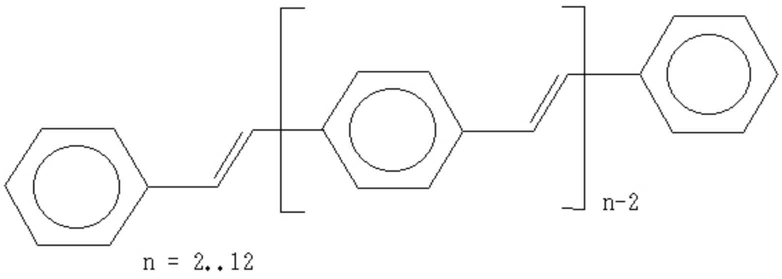

Figure 9 – PPV oligomers model.

Fourth group was composed by: β -carotene, chlorophyll, non-planar 12 units polyanilin oligomer (PAN12), streptomycin and taxol – an anti-tumoral drug (Figs. 10 up to 14).

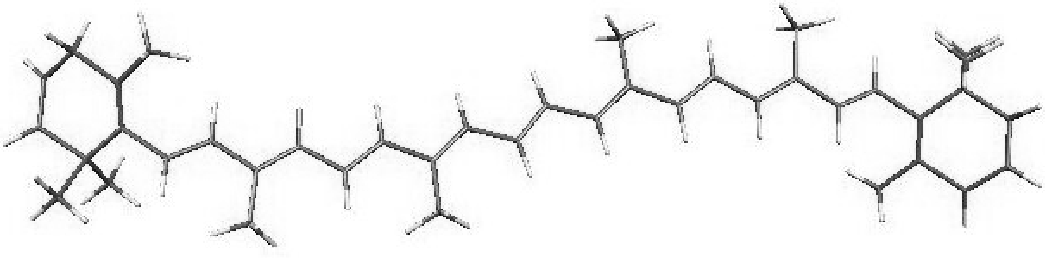

Figure 10 – β -carotene : one of the two forms of the dimer of vitamin A (40 Carbons, 52 Hydrogens)

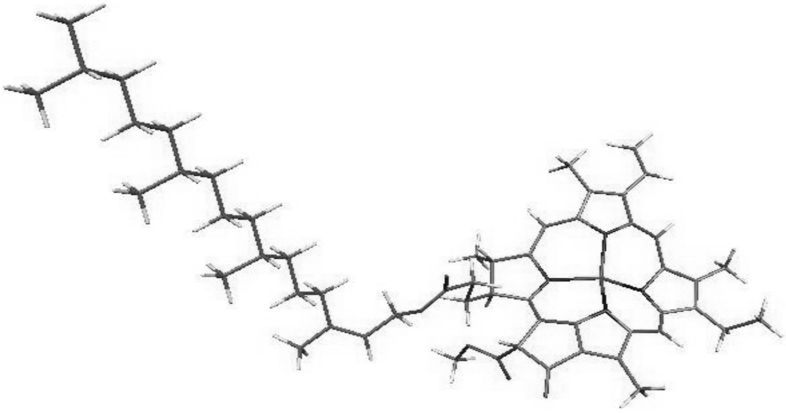

Figure 11 – Chlorophyll a: the molecule that absorbs sunlight and uses its energy to synthesise carbohydrates from CO2 and water (55 Carbons, 5 Oxygens, 1 Magnesium, 4 Nitrogens, 72 Hydrogens)

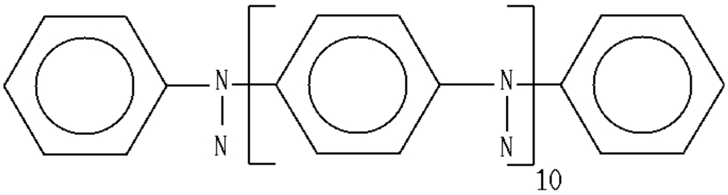

Figure 12 – PAN12: a non-planar 12 units polyanilin oligomer (72 Carbons, 11 Nitrogens and 61 Hydrogens)

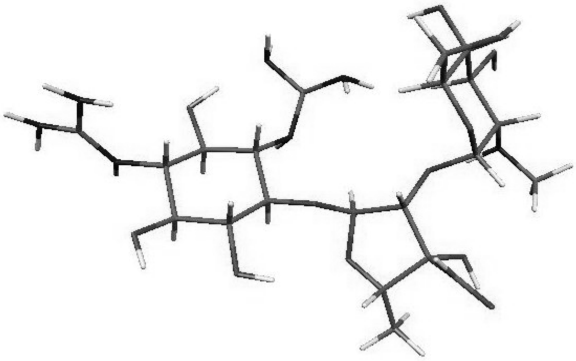

Figure 13 – Streptomycin: a antibiotic (21 Carbons, 12 Oxygens, 7 Nitrogens, 41 Hydrogens);

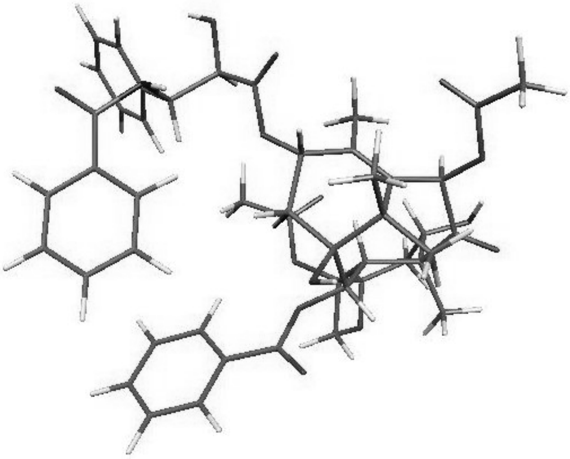

Figure 14 – Taxol: a anti-tumoral drug; (47 Carbons, 14 Oxygens, 1 Nitrogen, 51 Hydrogens);

Tables 2 a, 2 b, 2 c e 2 d show the results from equations 3.1, 3.2 and 3.3, when considering 6-31G basis set. Values presented in Table 2 allow to compare the real amount of 2e- integrals obtained from GAMESS execution (third column) to theoretical amount of them (last column). Theoretical amount of 2e- integrals was evaluated by means of direct multiplication of 2e- integrals peak amount (second column) by respective cutoff (sixth column). The cutoff uses the arithmetic average of the inertia moments and it is obtained according to description done in the previous section.

Table 2 (a)

Molecule	Integrals Amount (Peak)	Integrals Amount (Real)	Inertia Moments Average	Integrals Amount Index	Cutoff$_{3D}$ (eq. 3.3)	Integrals Amount (Theoretical)
C20	131220000	110959278	0.66	0.8456	0.8525	111862938
C24	272097792	208410000	0.9706	0.7659	0.7680	208984058
C26	374777442	273210000	1.1374	0.7290	0.7291	273247446
C32	859963392	546123744	1.7114	0.6351	0.6192	532500682
C36	1377495072	765225000	2.1683	0.5555	0.5514	759589474
C50	5125781250	1846036324	4.1746	0.3601	0.3652	1871907330
C60	10628820000	2871443565	5.9383	0.2702	0.2754	2927453593
C70	19691201250	3875575000	8.2007	0.1968	0.2040	4016438174
C80	33592320000	4935675000	10.8287	0.1469	0.1516	5093806765
tube9x0	53808401250	4784685000	15.6514	0.0889	0.0952	5123540728
tube5x5	82012500000	6374865000	19.2767	0.0777	0.0693	5685952011
tube10x0	82012500000	5480295000	19.7194	0.0668	0.0667	5474319381
tube11x0	120074501250	6276855000	24.5229	0.0523	0.0442	5308223325

Table 2 – List of distinct molecules comparing the Integrals Amount Real (empirical) with the Integrals Amount Theoretical, this obtained by means of modeling proposal. The results of cutoff (Eq. 3.3, 3.4 and 3.5) are also showed. Hartree-Fock results were obtained with a 6-31G basis set and a 10^{-10} hartrees integral screening threshold. The scale of the inertia-moment's averages was reduced dividing it by 10^6. Table 2 (a) groups molecules with tri-dimensional structure; Table 2 (b) binds molecules with planar structure, Table 2 (c) join linear ones and, finally, Table 2 (d) shows molecules with different geometry.

Table 2 (b)

Molecule	Integrals Amount (Peak)	Integrals Amount (Real)	Inertia Moments Average	Integrals Amount Index	Cutoff$_{planar}$ (eq. 3.4)	Integrals Amount (Theoretical)
C024H12	414720000	91891108	1.97	0.2216	0.2145	88976365
C054H18	9280941282	656138426	9.9010	0.0707	0.0650	603303957
C096H24	86474760192	2285019245	32.0117	0.0264	0.0226	1953694292
C150H30	288909830440	4307400000	63.3191	0.0149	0.0121	3490016083
ST 1	49406770	12405000	1.4198	0.2511	0.2556	12629446
ST 2	746602082	70680000	6.0850	0.0947	0.0976	72839839
ST 3	3707681328	187830000	11.3334	0.0507	0.0578	214330734
ST 4	11605565952	362790000	20.0194	0.0313	0.0347	403256711
ST 5	28170003480	588345000	32.0003	0.0209	0.0226	636643098
ST 6	58187567522	860745000	49.0567	0.0148	0.0153	887498257
ST 7	107501657770	1186680000	71.4531	0.0110	0.0108	1163878713
ST 8	183012500000	1552095000	100.7056	0.0085	0.0080	1458440458

Table 2 (c)

Molecule	Integrals Amount (Peak)	Integrals Amount (Real)	Inertia Moments Average	Integrals Amount Index	Cutoff$_{Linear}$ (eq. 3.5)	Integrals Amount (Theoretical)
PAH5	326094722	53407147	3.1	0.1638	0.1662	54194436
PAH6	625801442	77464180	4.9786	0.1238	0.1269	79440060
PAH7	1095962562	105564358	7.5165	0.0963	0.0967	106008265
PAH8	1791490082	139350678	10.6247	0.0778	0.0753	134885573
PAH9	2774976002	175805880	14.6980	0.0634	0.0587	162992453
PAH10	4116692322	216111579	19.7086	0.0525	0.0466	191807793
Ppv02	63281250	14740941	1.4130	0.2329	0.2310	14617302
Ppv03	374777442	41679123	5.3426	0.1112	0.1214	45516493
Ppv04	1278257922	79897863	13.4387	0.0625	0.0630	80489041
Ppv05	3264481602	130164851	27.1820	0.0399	0.0360	117597342
Ppv06	6973568802	191671635	48.0815	0.0275	0.0230	160044508
Ppv07	13195001250	264129359	77.6390	0.0200	0.0161	211908766
Ppv08	22867622082	347291656	117.3565	0.0152	0.0121	277597624
Ppv09	37079635842	440803634	168.8092	0.0119	0.0098	361853679
Ppv10	57068608482	544969890	233.2786	0.0095	0.0082	470288269
Ppv11	84221467362	659156710	312.4870	0.0078	0.0072	608549489
Ppv12	120074501250	783349636	407.8687	0.0065	0.0065	783228368

Table 2 (d)

Molecule	Integrals Amount (Peak)	Integrals Amount (Real)	Inertia Moments Average	Integrals Amount Index	Cutoff	Integrals Amount (Theoretical)
pan12	71283516990	1337184754	213.14	0.0188	0.00862	614294147
chlorophyl	36084933690	1597738503	43.6800	0.0443	0.02473	892289058
-carotene	5794045952	366675785	34.2000	0.0633	0.02998	173725839
streptomycin	23718420000	2129945016	13.8300	0.0898	0.11262	2671073415
taxol	4770886562	589268567	8.4100	0.1235	0.08937	426395108

Figures 15 to 18 show the values presented in the Table 2 in a graphical way. These results show that the modeling proposed in this work keeps values very close to those empirically obtained from algorithm execution. The errors observed in 3D, planar and linear groups were 3%, 7% and 8%, respectively. The differences verified with tube 5x5, tube 11x0 and C150H30 molecules were not considered significant. Their values are consistent and demonstrate that the curve of the modeled values is close to the empirical ones. The distinct-geometry molecular group, Table 2(d), present a major error rate, from 25% up to 50%. The irregular geometry observed in these molecules make difficult to match them in the proposed groups. However, the differences found here are significantly smaller than the observed in (SCHMIDT, 1993).

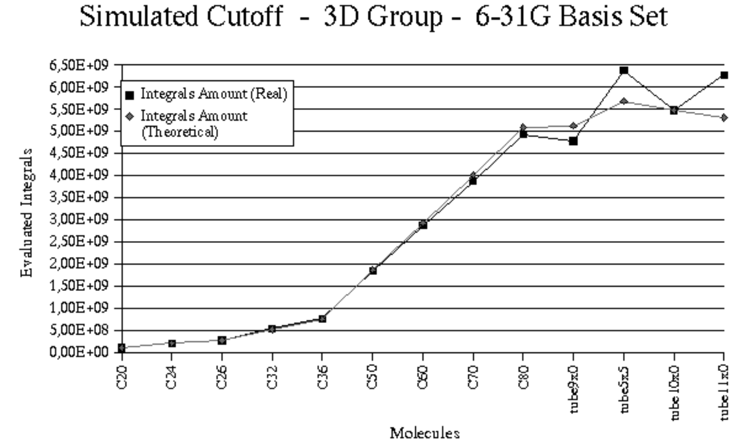

Figure 15 – Graph comparing real and theoretical 2e- Integral Amount for tri-dimensional molecular structures.

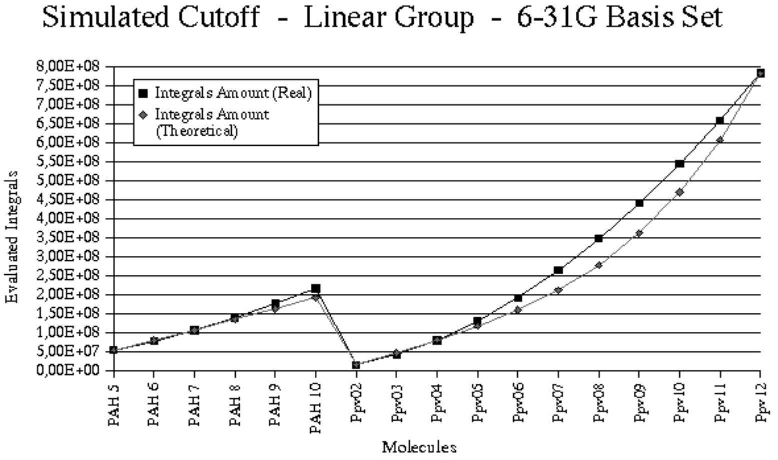

Figure 16 – Graph comparing real and theoretical 2e- Integral Amount for planar molecular structures.

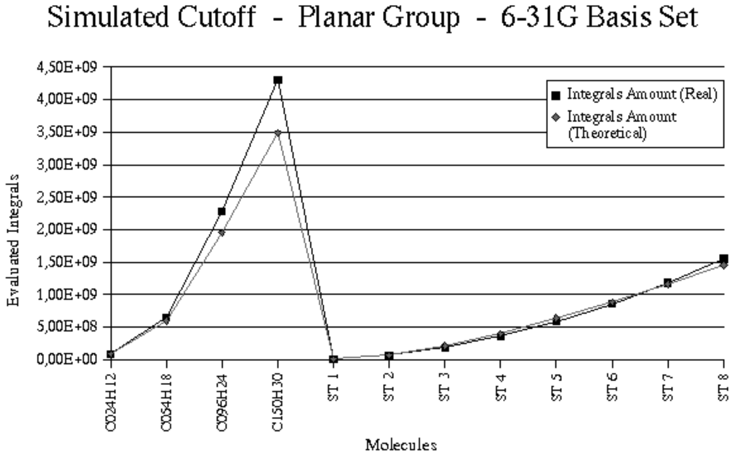

Figure 17 – Graph comparing real and theoretical 2e- Integral Amount for linear molecular structures.

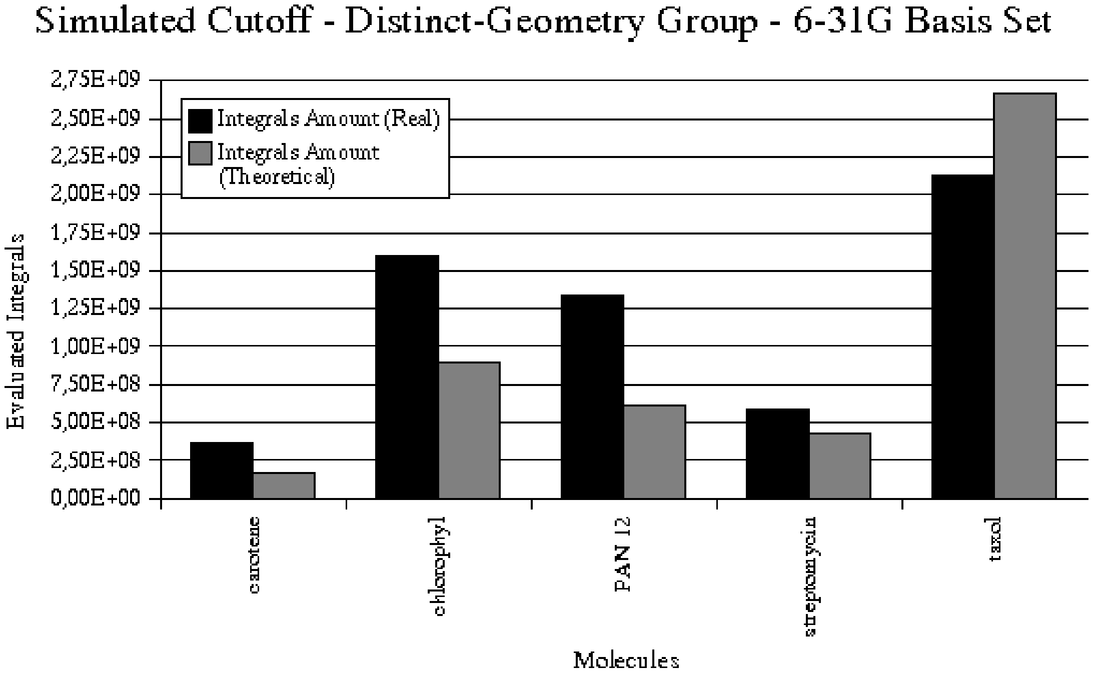

Figure 18 – Graph comparing real and theoretical 2e- Integral Amount for distinct-geometry group of molecular structures.

It can be observed in Fig. 19 that the modeling proposed here is consistent even when considering the STO-6G and 6-311G basis set. These results show a similar behavior to the 6-31 ones. Some fullerenes (C36 up to C80) cannot be executed with 6-311G basis set, since GAMESS presented an overflow error when summing the 2e- integrals amount. GAMESS uses 32 bits to represent the integrals amount and this upper threshold was exceeded.

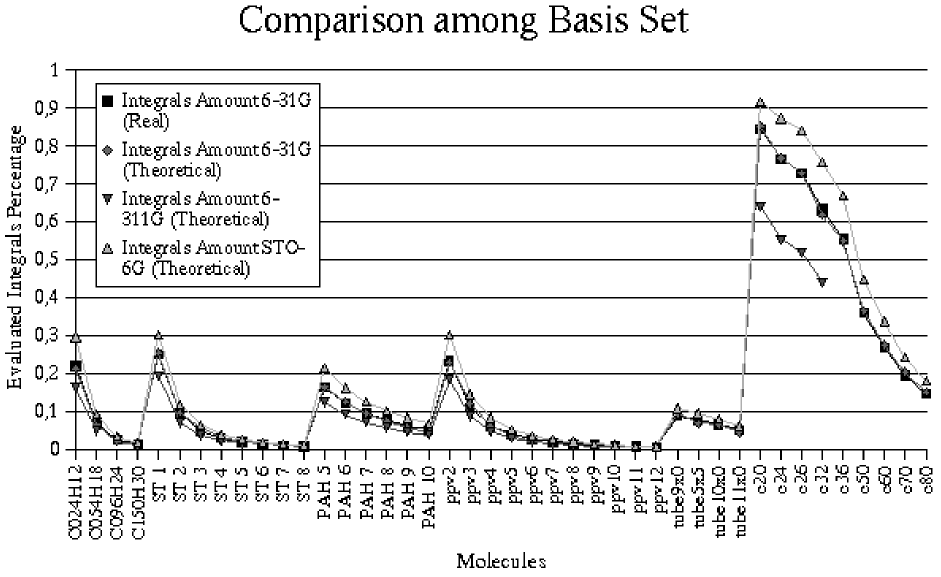

Figure 19 – Graph comparing Evaluated 2e- Integrals percentage among basis set.

Table 3 and Fig. 20 allow to compare the methodology proposed, considering both real and estimated exponent (α). Equation 3.1 was used as base for this comparison. Real α was obtained from empirical amount of integrals actually evaluated by GAMESS. Estimated α considered the amount of integrals predicted from cutoff modeling. The greatest differences found were 1.03% for regular molecular structures and 3.43% for the molecules inside fourth group.

Table 3 – Comparison of real and estimated α exponent.

Molecule	Real	Estimated	Difference
C20	3.96770	3.96927	0.03937
C24	3.95039	3.95090	0.01295
C26	3.94206	3.94208	0.00064
C32	3.91982	3.91536	0.11380
C36	3.89831	3.89703	0.03280
C50	3.83284	3.83511	0.05943
C60	3.79198	3.79505	0.08097
C70	3.74782	3.75336	0.14779
C80	3.70851	3.71330	0.12925
tube9x0	3.63864	3.64886	0.28080
tube5x5	3.62447	3.60766	0.46386
tube10x0	3.60224	3.60208	0.00445
tube11x0	3.57214	3.54784	0.68025
PAH 5	3.66622	3.66892	0.07364
PAH 6	3.62582	3.63033	0.12441
PAH 7	3.59115	3.59189	0.02042
PAH 8	3.56319	3.55762	0.15633
PAH 9	3.53675	3.52405	0.35926
PAH 10	3.51325	3.49354	0.56087
Ppv02	3.70923	3.70755	0.04532
Ppv03	3.59740	3.61354	0.44879
Ppv04	3.51883	3.52011	0.03636
Ppv05	3.46267	3.44574	0.48900
Ppv06	3.41901	3.38986	0.85260
Ppv07	3.38365	3.34893	1.02594
Ppv08	3.35412	3.31957	1.03008
Ppv09	3.32885	3.29897	0.89776
Ppv10	3.30699	3.28503	0.66402
Ppv11	3.28768	3.27594	0.35685
Ppv12	3.27044	3.27042	0.00069
C024H12	3.72503	3.71915	0.15789
C054H18	3.57662	3.56321	0.37509
C096H24	3.46689	3.44391	0.66296
C150H30	3.40907	3.37950	0.86728
ST 1	3.72074	3.72436	0.09738
ST 2	3.58111	3.58646	0.14936
ST 3	3.50523	3.52712	0.62461
ST 4	3.45111	3.46786	0.48534
ST 5	3.40802	3.42010	0.35423
ST 6	3.37265	3.37721	0.13512
ST 7	3.34406	3.34124	0.08445
ST 8	3.31888	3.30999	0.26778
PAN12	3.41246	3.29752	3.36829
chlorophyll	3.52748	3.43918	2.50332
β-carotene	3.55046	3.42880	3.42669
taxol	3.62876	3.66363	0.96093
streptomycin	3.65666	3.60355	1.45245

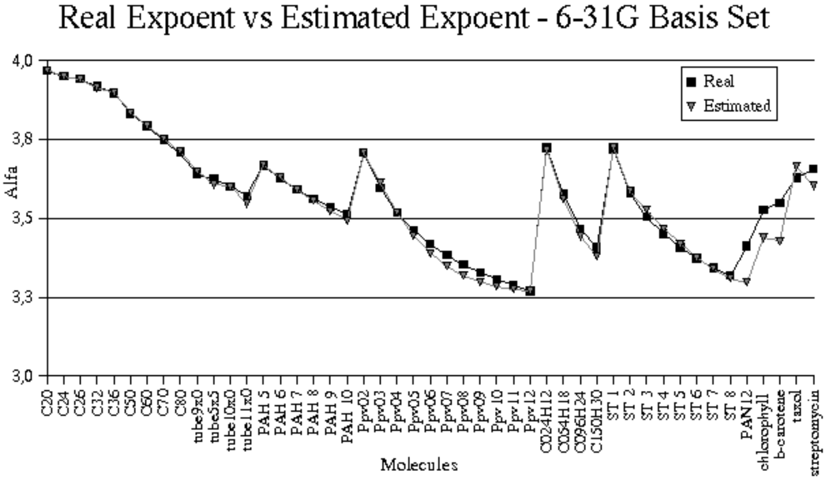

Figure 20 – Graph comparing real and estimated α exponent.

These results show that is possible to estimate, with better precision, the complexity of the energy evaluation with the Hartree-Fock algorithm. This study is useful to the user because it makes possible to him to appraise formerly the computational cost generated by HF procedure. It considers as example the sequential execution of tube10x0 molecule, storing temporally in disk the 2e- integrals evaluated in HF procedure. The modeling proposal here allow to estimate that all 5.74 x 10^9 .2e- integrals will require 81.57 Gbytes, in order to store in disk the integrals and their labels (8+8 bytes). This result can be used directly, for example, to determine the execution viability with Conventional SCF – CSCF algorithm, which stores the integrals in disk.

The estimative of the 2e- integrals amount existent in GAMESS manual (SCHMIDT, 1993) is the nearest to this work. The GAMESS estimative does not consider the 2e- integrals cutoff and gives to the final user a maximum limit to integrals amount. The estimative provided by GAMESS for the early example would be ~82.01 x 10^9 integrals and ~1.19Tbytes stored in disk. The use of ~N^2 is far from real amount of 2e- integrals in this case, since it would result ~8.1 x10^5 integrals and 12.36Mbytes stored in disk.

If there is a trustful estimative of the 2e- integrals amount, it will be possible estimate the time necessary to execute the HF procedure too. The time estimative does not belong to scope of this work. However, a future analytic modeling can instantiate the proportionally of 2e-integrals amount and the time to execute the HF procedure.

5 Concluding Remarks and Future Works

This paper presented a methodology to estimate real 2e- integrals amount in HF procedure, when different kind of molecules are executed. This methodology is based in molecular geometry and uses as metric the inertia moment mean.

Prime studies were done using 6-31G basis set. This basis set was chosen because it is an intermediary basis set and due to its common use. STO-6G and 6-311G were also used to demonstrate the behavior of proposed methodology in different situations.

Results obtained attest that proposed methodology is consistent and allow estimating 2e-integrals cutoff for unknown geometry molecules. Errors observed in modeling were 3%, 7% e 8% respectively to 3D, planar and linear groups.

Errors found in distinct geometry molecular group are caused by the lack of regular geometry, primordially. However, the fact of this work estimates 2e- integrals cutoff based on 3D, planar and linear geometries does not hinder its use in other situations. Estimate remains consistent for geometry distinct molecules, hence differences found are significantly smaller than that observed in (SCHMIDT, 1993).

Future works on this subject are mainly directed to determine the complexity existent in their algorithms. This work will be useful to the final user that will be able to predict the necessary time to execute their simulations and to determine the disk demand too.

It is being developed a methodology to determine the scalability of HF procedure on a Beowulf Cluster. The 2e- integrals cutoff methodology proposed here is being used successfully.

Another point, not treated in this work, is the automating of molecular geometry choice. The modeling proposed here requires that user indicates which is the most suitable model of three groups for its molecule.

Acknowledgements

The authors thank the CNPq for the financial support given to this work.

References

AHLRICHS, R.; et al. Electronic Structure Calculations on Workstation Computers: The Program System Turbomole. **Chemical Physics Letters**, v.162, p.165, 1989.

AIKENS, C.M.; GORDON, M.S. Parallel Unrestricted MP2 Analytic Gradients Using the Distributed Data Interface. **Journal of Physical Chemistry,** *v.108, p.3103, 2004.*

ALEXEEV, Y.; KENDALL, R.A.; GORDON, M.S. The distributed data SCF. **Computer Physics Communication,** v.143, p.69, 2002.

ALMLÖF, J.; FAEGRI JR, K.; KORSELL, K. **Principles for a direct SCF approach to LICAO-MO ab-initio Calculations**. **Journal of Computacional Chemistry,** v.3, p.385, 1982.

BOLDING, B.; BALDRIDGE, K. Multithreaded shared memory parallel implementation of the electronic structure code GAMESS. **Computer Physics Communication,** v.128, p.55, 2000.

CHALLACOMBE, M.; SCHWEGLER, E. Linear scaling computation of the fock matrix. **Journal of Chemical Physics,** v.106, p.5526, 1997.

CHALLACOMBE, M. Linear scaling computation of the fock matrix. V. Hierarchical cubature for numerical integration of the exchange-correlation matrix. **Journal of Chemical Physics,** v.113, p.10037, 2000.

CHOI, C.H.; KOREAN, B. **Chemistry Society,** v.24, p.733, 2003.

FAMULARI, A.; et al. Hartree-fock limit properties of the water dimer in absence of BSSE. **Chemiscal Physics,** *v*.232, p. 275, 1998.

FEDOROV, D.M.; et al. A new hierarchical parallelization scheme: Generalized Distributed Data Interface (GDDI), and an application to the Fragment Molecular Orbital Method (FMO). **Journal of Computational Chemistry**, v.25, p.872, 2004.

FRISCH, M. J.; et al. Gaussian, Inc., Wallingford CT. **Gaussian 03, Revision C.02**, 2004.

GAN, C.K.; CHALLACOMBE, M. Linear scaling computation of the fock matrix. vi. data parallel computation of the exchange-correlation matrix. **Journal of Chemical Physics**, v.118, p.9128, 2003.

GAN, Z.; et al. The parallel implementation of a full configuration interaction program. **Journal of Chemical Physics.** v.119, p.47, 2003.

GLAESEMANN, K.R.; GORDON, M.K. Investigation of a grid-free Density Functional Theory (DFT) Approach. **Journal of Chemical Physics**, v.108, p.9959, 1998.

GRACE software: <http://plasma-gate.weizmann.ac.il/Grace>. Last access: July 17, 2009.

HELGAKER, T.; et al. **Dalton, a molecular electronic structure program, Release 1.2**, 2001.

KENDALL, R.A. et al. High performance computational chemistry: an overview of nwchem a distributed parallel application. **Computer Physics Communication,** v.128, p.260, 2000.

SCHMIDT, M.W.; et al. **General atomic and molecular electronic structure system. Journal of Computational Chemistry,** v.14, p.1347, 1993.

SCHWEGLER, E.; CHALLACOMBE, M. Linear scaling computation of the hartree–fock exchange matrix. **Journal of Chemical Physics,** v.105, p.2726, 1996.

SCHWEGLER, E.; CHALLACOMBE, M.; HEAD-GORDON, M. Linear scaling computation of the fock matrix. II. Rigorous bounds on exchange integrals and incremental fock build. **Journal of Chemical Physics**, v.106, p.9708, 1997.

SCHWEGLER, E.; CHALLACOMBE, M. Linear scaling computation of the fock matrix. IV. Multipole accelerated formation of the exchange matrix. **Journal of Chemical Physics,** v.111 p.6223, 1999.

SCHWEGLER, E.; CHALLACOMBE, M. Linear scaling computation of the fock matrix. III. Formation of the exchange matrix with permutational symmetry. **Theoretical Chemistry Accounts,** v.104, p.344, 2000.

STROUT, D.L.; , SCUSERIA, G.E. A quantitative study of the scaling properties of the hartree-fock method. **Journal of Chemical Physics,** v.102, p.8448, 1995.

TRUHLAR, D.G. Perspective on "principles for a direct SCF approach to LCAO-MO ab-initio calculations". **Theoretical Chemistry Accounts,** v.103, p.349, 2000.

TYMCZAK, C.J.; CHALLACOMBE, M. Linear scaling computation of the fock matrix. VII. Periodic density functional theory at the Γ point. Journal of Chemical Physics, v.122 , p.134102, 2005a.

TYMCZAK, C.J. et al. Linear scaling computation of the fock matrix. VIII. Periodic boundaries for exact exchange at the Γ point. **Journal of Chemical Physics**, v.122, p.124105, 2005b.

SPARTAN, Wavefunction, Inc., 18401 Von Karman Ave., No. 370, Irvine, CA 92715.